Complex Genomes of Early Nucleocytoviruses Revealed by Ancient Origins of Viral Aminoacyl-tRNA Synthetases

Soichiro Kijima [1,2,†] Hiroyuki Hikida [1,†] Tom O. Delmont [3] Morgan Gaïa [3] Hiroyuki Ogata [1,*]

[1]Chemical Life Science, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan

[2]School of Life Science and Technology, Tokyo Institute of Technology, Meguro, Tokyo 152-8550, Japan

[3]Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ. Evry, Université Paris-Saclay, 91057 Evry, France

[†]These authors contributed equally to this work.

***Corresponding author:** E-mail: ogata@kuicr.kyoto-u-ac.jp.

Associate editor: Thomas Leitner

Abstract

Aminoacyl-tRNA synthetases (aaRSs), also known as tRNA ligases, are essential enzymes in translation. Owing to their functional essentiality, these enzymes are conserved in all domains of life and used as informative markers to trace the evolutionary history of cellular organisms. Unlike cellular organisms, viruses generally lack aaRSs because of their obligate parasitic nature, but several large and giant DNA viruses in the phylum *Nucleocytoviricota* encode aaRSs in their genomes. The discovery of viral aaRSs led to the idea that the phylogenetic analysis of aaRSs can shed light on ancient viral evolution. However, conflicting results have been reported from previous phylogenetic studies: one posited that nucleocytoviruses recently acquired their aaRSs from their host eukaryotes, while another hypothesized that the viral aaRSs have ancient origins. Here, we investigated 4,168 nucleocytovirus genomes, including metagenome-assembled genomes (MAGs) derived from large-scale metagenomic studies. In total, we identified 780 viral aaRS sequences in 273 viral genomes. We generated and examined phylogenetic trees of these aaRSs with a large set of cellular sequences to trace evolutionary relationships between viral and cellular aaRSs. The analyses suggest that the origins of some viral aaRSs predate the last common eukaryotic ancestor. Inside viral aaRS clades, we identify intricate evolutionary trajectories of viral aaRSs with horizontal transfers, losses, and displacements. Overall, these results suggest that ancestral nucleocytoviruses already developed complex genomes with an expanded set of aaRSs in the proto-eukaryotic era.

***Key words:* *Nucleocytoviricota*, aminoacyl-tRNA synthetase, ancestral origin.**

Introduction

Nucleocytoviricota is a phylum of dsDNA viruses (nucleocytoviruses), formerly known as nucleocytoplasmic large DNA viruses, which infect diverse eukaryotes from microeukaryotes to animals (Iyer et al. 2006; Guglielmini et al. 2019; Aylward et al. 2021). These viruses have large genomes that can exceed 1 Mb, with over 1,000 genes (Raoult et al. 2004; Philippe et al. 2013). Their genomes encode genes that were previously considered to be exclusive to cellular organisms, such as those related to energy production, chromatin remodeling, and translation (Raoult et al. 2004; Boyer et al. 2009; Schvarcz and Steward 2018; Yoshikawa et al. 2019; Blanc-Mathieu et al. 2021). Although these cellular hallmark genes in viral genomes were speculated to be derived from cellular organisms, their origin and evolution remain largely elusive. Some proposed the fourth-domain hypothesis based on deep branches of these virus-encoded cellular hallmark genes independent from those encoded in the three established domains of life (Raoult et al. 2004; Claverie 2006; Boyer et al. 2010; Boughalmi et al. 2013). Under this rather provocative hypothesis, nucleocytoviruses originated from ancient and extinct cellular organisms with a complete set of cellular genes and experienced reductive evolution, thereby still retaining many cellular hallmark genes. Meanwhile, others proposed an alternative model in which a small viral ancestor accumulated genes from cellular organisms, like a "gene robber," by horizontal gene transfer (HGT) or de novo gene creation (Williams et al. 2011; Yutin et al. 2014; Moreira and López-García 2015; Forterre and Gaïa 2016; Legendre et al. 2018). All scenarios, however, suppose that nucleocytoviruses and their host organisms underwent long-lasting interactions. The phylogeny of genes conserved within *Nucleocytoviricota* supported the occurrence of such interactions by demonstrating that ancestral nucleocytoviruses had already

Received: March 05, 2024. **Revised:** June 27, 2024. **Accepted:** July 15, 2024

appeared in the proto-eukaryotic era, a period before the last eukaryotic common ancestor (LECA) during which the complex and unique traits of eukaryotic cells were developed (Koonin and Yutin 2010; Mihara et al. 2018; Guglielmini et al. 2019). Tight evolutionary associations between eukaryotes and nucleocytoviruses are also supported by HGTs, some of which date back to the proto-eukaryotic period (e.g. two largest subunits of the viral DNA-dependent RNA polymerase, actin-related protein, and topoisomerase type IIA) (Guglielmini et al. 2019; Da Cunha et al. 2022; Guglielmini et al. 2022; Irwin et al. 2022).

Aminoacyl-tRNA synthetases (aaRSs) are among the cellular hallmark enzymes encoded by nucleocytoviruses whose evolution has been controversial (Abrahão et al. 2017). These enzymes have a central role in decoding genetic codes and are conserved in all three domains of life, but small viruses rarely encode them. A nucleocytovirus, Acanthamoeba polyphaga mimivirus (APMV), is the first virus found to encode aaRSs (ArgRS, CysRS, MetRS, and TyrRS) (Raoult et al. 2004). Isolation of its relative (*Megavirus chilensis*) further expanded the repertoire of viral aaRSs from four to seven (IleRS, TrpRS, and AsnRS, in addition to the set encoded by APMV) (Arslan et al. 2011). Discoveries of nucleocytoviruses with an increasing number of aaRSs raised expectations that their ancestor had a complete set of 20 aaRSs, which would support the reductive evolution scenario (Abrahão et al. 2017). However, phylogenetic analysis of these aaRSs suggested their recent origins by gradual accumulation through HGT from eukaryotic hosts based on polyphyly and tree topologies suggesting eukaryotic origins for some aaRSs (Yutin et al. 2014). Recently, nucleocytoviruses with more enriched sets of aaRSs were discovered, but the uncertainty remained. Metagenome-derived klosneuvirus was found to encode 19 aaRSs, and their phylogenies supported scenarios of recent origins for most of these genes; 17 out of 19 genes appeared to have been horizontally transferred from different eukaryotes (Schulz et al. 2017). In contrast, a study of tupanviruses, which were isolated from extreme environments and found to encode the complete set of 20 aaRSs, claimed that the origins of tupanvirus aaRSs are not necessarily explained by recent HGT from eukaryotes (Abrahão et al. 2018).

aaRSs are conserved in all domains of life, and therefore, they are useful markers to resolve ancient evolution of cellular organisms (Woese et al. 2000). However, unlike other marker genes to resolve the tree of life, evolutionary trajectories of aaRSs are complex with frequent HGT and subsequent displacements (Doolittle and Handy 1998; Wolf et al. 1999). Eukaryotic aaRSs show particularly complex evolutionary histories due to multiple sets of homologs that function in different translational compartments: cytosol, mitochondria, and chloroplasts (plastids). These characteristics might have hampered an unequivocal interpretation of phylogenetic trees of viral aaRSs. Furthermore, previous phylogenetic analyses of viral aaRSs used a small number of sequences from nucleocytoviruses and incomprehensive taxon sampling of cellular organisms (Yutin et al. 2014; Schulz et al. 2017; Abrahão et al. 2018). Thus, the ability to resolve their evolutionary histories could have been limited by computational artifacts (e.g. long-branch attraction [LBA]). In the present study, we revisit the origins of nucleocytovirus aaRSs by using recently reconstructed MAGs (Schulz et al. 2020; Moniruzzaman, et al. 2020a; Gaïa et al. 2023) as well as cellular aaRSs from a broad range of taxa (Furukawa et al. 2017). By reconstructing large-scale phylogenetic trees of aaRSs, we reliably identified the origins of some viral aaRSs, which highlight the ancient genome complexity of nucleocytoviruses.

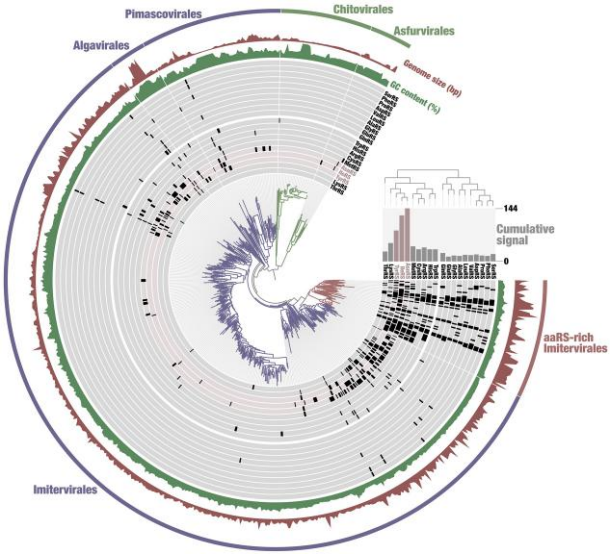

Fig. 1. A phylogenetic tree of nucleocytovirus species based on the concatenated sequences of eight-core genes. The layers surrounding the tree represent the distribution of aaRSs. The substitution model was Q.pfam + F + R10. *Asfuvirales* and *Chitovirales* were used as outgroups following the method reported by Aylward et al. (2021).

Results

Identification of Nucleocytovirus aaRSs

From 224 reference nucleocytovirus genomes and 3,578 nonredundant nucleocytovirus MAGs, we identified 780 aaRS sequences in 273 genomes (supplementary table S1, Supplementary Material online). PheRS, GlyRS, and LysRS have multiple types of enzymes, but those detected in nucleocytoviruses were all related to a single type: the cellular PheRS α subunit, GlyRS-1, and LysRS-II, respectively. The most abundant aaRSs were associated with AT-rich codons, namely, 144 AsnRSs (codons: AAU/C), 125 IleRSs (AUU/C/A), and 84 TyrRSs (UAU/C), representing 18%, 16%, and 11% of the total aaRSs, respectively (supplementary fig. S1, Supplementary Material online). Out of 780 aaRSs, 730 (93.6%) were encoded by viruses in the order *Imitervirales* (Fig. 1; supplementary table S1, Supplementary Material online). Within this order, an internal monophyletic clade consisting of 64 genomes harbored 398 aaRSs, which is over half of the aaRSs encoded in this order. This clade included nucleocytoviruses known to encode an expanded set of aaRSs, such as mimiviruses,

Table 1 Evolutionary scenarios of viral aaRSs

m	aaRS	Clade/branch[a]	Scenario[b]	Support[c]	
A	AlaRS	...	Recent Euk to V	Topology and UBF/SH-aLRT	...
C	CysRS	...	Ancient Euk&V	Topology and UBF/SH-aLRT	...
D	AspRS	Clade I	Proto Euk&V	Topology and UBF/SH-aLRT	...
		Clade II	Other Euk&V	Topology and UBF/SH-aLRT	Tupanviruses and one MAG
E	GluRS	...	Proto Euk&V	Topology and UBF/SH-aLRT	...
F	PheRS	...	Proto Euk&V	Topology and UBF/SH-aLRT	Putatively from organelles
G	GlyRS	Clade I	Proto Euk&V	Topology and UBF/SH-aLRT	...
		Branch I	Prok to V	Topology and UBF/SH-aLRT	Orpheovirus
H	HisRS	Clade I	Other Euk&V	Topology and UBF/SH-aLRT	...
		Clade II	V to Euk	Topology	...
		Clade III	Prok to V	Topology and UBF/SH-aLRT	...
I	IleRS	...	Proto Euk&V	Topology and UBF/SH-aLRT	...
K	LysRS	...	Proto Euk&V	Topology and TBE	...
L	LeuRS	...	Proto Euk&V	Topology and TBE	...
M	MetRS	Clade I	Ancient Euk&V	Topology and UBF/SH-aLRT	...
		Clade II	Ancient Euk&V	Topology and UBF/SH-aLRT	...
		...	Recent Euk to V	Topology	Klosneuvirus, hokovirus, and two MAGs
N	AsnRS	Clade I	Ancient Euk&V	Topology and UBF/SH-aLRT	...
		Clade II	Prok to V	Topology	...
P	ProRS	Clade I	Other Euk&V	Topology	...
		Clade II	Other Euk&V	Topology	...
Q	GlnRS	...	Other Euk&V	Topology	...
R	ArgRS	Clade I	Recent Euk to V	Topology	Orpheovirus and one MAG
		Clade II	Other Euk&V	Topology	...
S	SerRS	Clade I	Other Euk&V	Topology and UBF/SH-aLRT	...
		Clade II	Recent Euk to V	Topology and UBF/SH-aLRT	Tupanviruses and one MAG
		Branch I	Recent Euk to V	Topology and UBF/SH-aLRT	Orpheovirus
T	ThrRS	Clade I	Ancient Euk&V	Topology	...
		Clade II	Recent Euk to V	Topology	Tupanviruses and one MAG
V	ValRS	Clade I	Prok to V	Topology and UBF/SH-aLRT	Tupanviruses
		Clade II	Prok to V	Topology	...
W	TrpRS	Clade I	Other Euk&V	Topology and UBF/SH-aLRT	...
		Clade II	Other Euk&V	Topology and UBF/SH-aLRT	...
		Branch I	Recent Euk to V	Topology and UBF/SH-aLRT	Pandoravirus salinus
Y	TyrRS	Clade I	V to Euk	Topology and UBF/SH-aLRT	Encompassing a eukaryotic clade
		Clade II	V to Euk	Topology and UBF/SH-aLRT	Encompassing a eukaryotic clade

[a]Singleton of a MAG or clades consisting two MAGs were not considered.
[b]Euk: eukaryotes; Prok: prokaryotes; V: viruses.
[c]AA: amino acid; UBF: ultrafast bootstrap; SH-aLRT: Shimodaira–Hasegawa approximate likelihood ratio test; TBE: transfer bootstrap expectation.

tupanviruses, and klosneuviruses (Raoult et al. 2004; Schulz et al. 2017; Abrahão et al. 2018).

Evolution of Nucleocytovirus aaRSs

To elucidate the evolutionary histories of nucleocytovirus aaRSs, we constructed their phylogenetic trees together with a comprehensive set of cellular sequences from a previous study (Furukawa et al. 2017). This data set covers wide taxonomic groups: 64 eukaryotic species from 26 phyla, 142 bacterial species from 54 phyla, and 76 archaeal species from 11 phyla. The eukaryotic sequences encompassed both cytosolic and organellar sequences. Following the previous study (Furukawa et al. 2017), the root of each phylogenetic tree was decided at a branch separating major bacterial and archaeal clades. We estimated their evolutionary histories based on tree topology and their statistical support: ultrafast bootstrap (UFB > 95%) and Shimodaira–Hasegawa approximate likelihood ratio test (SH-aLRT > 80%). We further constructed subsection trees for some aaRSs to improve the resolution of tree topologies. We assigned an evolutionary scenario for each nucleocytovirus aaRS clade, as shown in Table 1. The scenarios were mainly separated into six categories: (i) ancient HGT between proto-eukaryotes and nucleocytoviruses (proto-Euk&V); (ii) ancient HGT between eukaryotes and nucleocytoviruses, but not necessarily dating back to proto-eukaryotes (ancient Euk&V); (iii) recent HGT from eukaryotes to nucleocytoviruses (recent Euk&V); (iv) HGT from nucleocytoviruses to eukaryotes (V to Euk); (v) HGT between eukaryotes and nucleocytoviruses for which the timing of transfer is less clear (other Euk&V); and (vi) HGT between prokaryotes and nucleocytoviruses (Prok&V). We describe details of selected notable cases below, while others are presented in the Supplementary Material.

Nucleocytoviral genes often deeply branched out in the reconstructed phylogenetic trees, which can be attributable to LBA in the inference of their evolutionary trajectory. Although it is generally difficult to assess the effect of LBA, we added two extra tests to augment the information on the stability/instability of the evolutionary scenario for the cases of ancient HGT between proto-eukaryotes

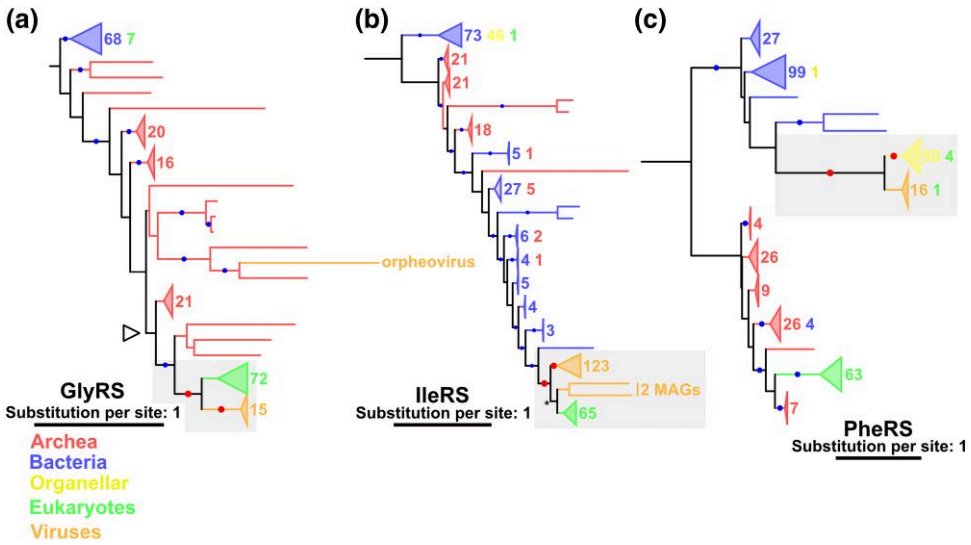

Fig. 2. Phylogenetic trees of a) GlyRS, b) IleRS, and c) PheRS. Red and blue dots indicate significant statistical support (≥80% SH-aLRT and ≥95% UFB). The numbers next to the collapsed clades and their colors indicate the number of sequences and their sources, respectively. Red dots and gray area indicate the nodes and the phylogenetic relationships described in the main text, respectively. a) An empty arrowhead indicates the node from which subsection trees were constructed. b) An asterisk indicates the node supported by UFB = 94% and SH-aLRT = 96.8%. The substitution models were a) LG + R9, b) Q.pfam + F + R10, and c) LG + R9. The root was decided following the work of Furukawa et al. (2017).

and nucleocytoviruses (proto-Euk&V). Specifically, we built additional phylogenetic trees using a more stringent trimming strategy of multiple sequence alignments (see Materials and Methods). We also generated constrained trees (3 topologies where viral sequences form a sister clade of the eukaryotic clade [original topologies] and 12 topologies where viral clade is within the eukaryotic clade [alternative topologies]) and calculated *P* values of each tropology by the approximately unbiased (AU) test (Shimodaira 2002).

Ancient HGT between Proto-eukaryotes and Nucleocytoviruses (Proto-Euk&V)

We identified HGTs between nucleocytoviruses and proto-eukaryotes in GlyRS, IleRS, AspRS, GluRS, LysRS, and LeuRS (Fig. 2; supplementary fig. S2, Supplementary Material online). We describe details of GlyRS and IleRS below and AspRS, GluRS, LysRS, and LeuRS in the Supplementary Material.

In the tree of GlyRS, three domains were largely separated, with a major eukaryotic clade branching from within the archaea domain (Fig. 2a). Two small eukaryotic clades were also found in bacterial clades, but species encoding these GlyRSs were also represented in the major eukaryotic clade. Therefore, the small eukaryotic clades likely consist of organellar GlyRSs derived from bacteria, while the root of the major eukaryotic clade corresponds to LECA. A majority of nucleocytovirus GlyRS sequences formed the sister clade to the major eukaryotic clade. Grouping of the viral GlyRSs with eukaryotic ones and the monophyly of the viral GlyRSs were statistically supported. However, the monophyly of the eukaryotic clade was not supported. We hence generated a subsection tree around the viral and eukaryotic GlyRSs, which supported the monophyly of the eukaryotic GlyRSs (supplementary fig. S3a, Supplementary Material online). With additional trimming strategies, the subsection tree kept the topology with statistical supports (supplementary fig. S4a, Supplementary Material online). These results indicate an ancient HGT for GlyRS between nucleocytoviruses and proto-eukaryotes.

A similar scenario was inferred for IleRS. In the IleRS tree (Fig. 2b), a monophyletic group of all viral sequences and most eukaryotic sequences except one in a bacterial clade was confirmed. Furthermore, a large majority of viral sequences formed a supported monophyletic group, while the monophyly of the clade containing eukaryotic sequences was not supported, which hampers the unambiguous placement of LECA. However, the clade of the eukaryotic sequence and two viral MAGs sister to the eukaryotic group had relatively high support values (UFB = 94%, SH-aLRT = 96.8%), although the UFB value did not meet the threshold that we applied (UFB > 95%). This led us to assume that the root of this clade corresponds to the time of LECA or earlier. A tree with a stringent trimming criterion also reproduced statistically supported separation between eukaryote and viral clades (supplementary fig. S4b, Supplementary Material online). Therefore, we conclude that viral IleRSs experienced an ancient HGT between nucleocytoviruses and proto-eukaryotes.

We also identified a putative and ancient HGT between organelles and nucleocytoviruses for PheRS. In the PheRS tree, bacterial and archaeal sequences are largely separated (Fig. 2c). Eukaryotic (nuclear) PheRSs were grouped with archaeal sequences, while nucleocytovirus PheRSs were grouped with bacterial and organellar sequences, indicating different origins for these PheRSs. This putative origin of viral PheRSs implies that their biochemical properties are also similar to mitochondrial PheRSs (Supplementary Material).

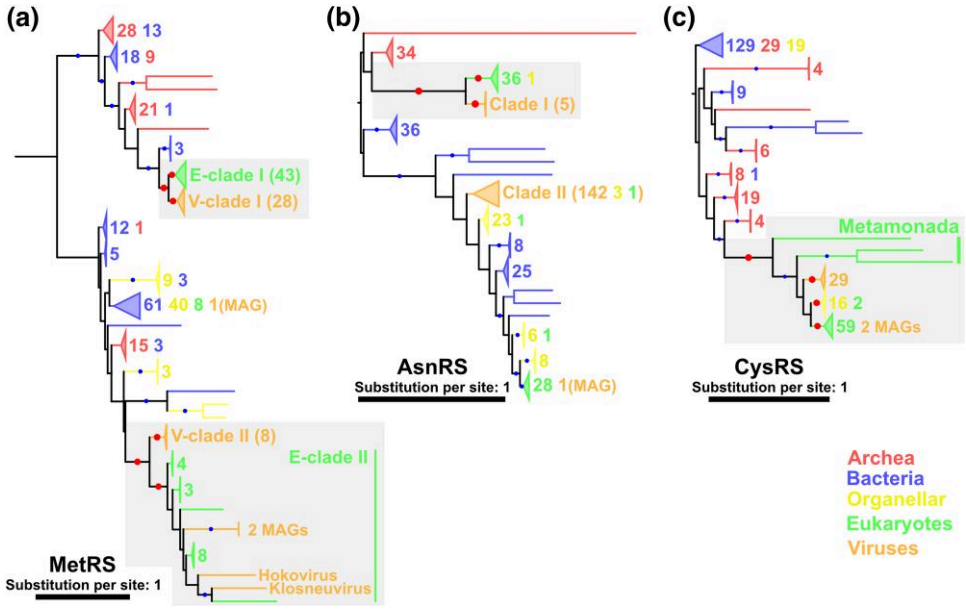

Fig. 3. Phylogenetic trees of a) MetRA, b) AsnRS, and c) CysRS. Red and blue dots indicate significant statistical support (≥80% SH-aLRT and ≥95% UFB). The numbers next to the collapsed clades and their colors indicate the number of sequences and their sources, respectively. Some clades are named, and the numbers in brackets indicate the number of sequences included within them. Red dots and gray area indicate the nodes and the phylogenetic relationships described in the main text, respectively. The substitution models were a) LG + R9, b) Q.pfam + F + R10, and c) LG + R9. The root was decided following the work of Furukawa et al. (2017).

This organellar clade included all organellar sequences except one and covered a broad range of eukaryotic taxa, suggesting that its root corresponds to the time of LECA. The statistically supported monophyly of organellar sequences and the grouping of the organellar and viral sequences suggest an ancient HGT between viruses and proto-eukaryotes (or organelles), which occurred before LECA. However, a phylogenetic tree with a trimmed alignment failed to reproduce the topology, and thus, we could not firmly establish their origin before LECA (supplementary fig. S4c, Supplementary Material online).

To support the robustness of the scenario, we performed AU test with constrained trees with 3 original topologies where viral sequences formed a sister clade of the eukaryotic clade and 12 alternative topologies where viral sequences were within the eukaryotic clade (supplementary table S2, Supplementary Material online). We failed to reject multiple alternative topologies for the seven aaRSs described here. However, *P* values of these alternative topologies were generally lower than those of original topologies except for LysRS (supplementary table S3, Supplementary Material online). Although we could not exclude alternative topologies with the AU test, the statistics are in favor of the scenario of ancient HGT between nucleocytoviruses and proto-eukaryotes.

Ancient HGT between Eukaryotes and Nucleocytoviruses (Ancient Euk&V)

We identified other cases of ancient HGTs between eukaryotes and nucleocytoviruses. In these cases, the HGT events were inferred to have occurred before the radiation of several major lineages of eukaryotes, although it was unclear whether these events date back to the proto-eukaryotic era (i.e. before LECA). The trees of MetRS, CysRS, and AsnRS suggested this scenario. A similar case was observed for ThrRS, although the scenario was not statistically supported for it (supplementary fig. S6, Supplementary Material online). We here describe the cases for MetRS and CysRS.

The MetRS tree contains two phylogenetically distant large eukaryotic clades (E-clades I and II) (Fig. 3a); one of these clades (E-clade II) contains four viral sequences. Each of the clades is sister to a clade consisting of viral MetRSs (V-clades I and II). In both clades, the monophyly of eukaryotes, the monophyly of nucleocytoviruses, and the grouping of these two adjust spacing were statistically supported, indicating the occurrence of HGTs between eukaryotes and nucleocytoviruses before the divergence of the respective eukaryotic clades. Members of the two eukaryotic clades did not overlap: E-clade I mainly included species of the Amorphea and Archaeplastida supergroups and E-clade II included members of the SAR supergroup. This phylogenetic distribution indicates that eukaryotes acquired their MetRS twice. As the two eukaryotic clades consist of different supergroups, it was unclear whether the roots of these eukaryotic clades correspond to LECA. Nevertheless, the HGTs between eukaryotes and viruses occurred before the divergence of each eukaryotic clade. Thus, we classified the HGTs of MetRS between E-clades and V-clades as ancient events. A clade of viral AsnRSs (clade I) also suggested a similar evolutionary history (Fig. 3b; Supplementary Material).

The CysRS tree depicts another scenario (Fig. 3c). This tree has a statistically supported clade consisting of

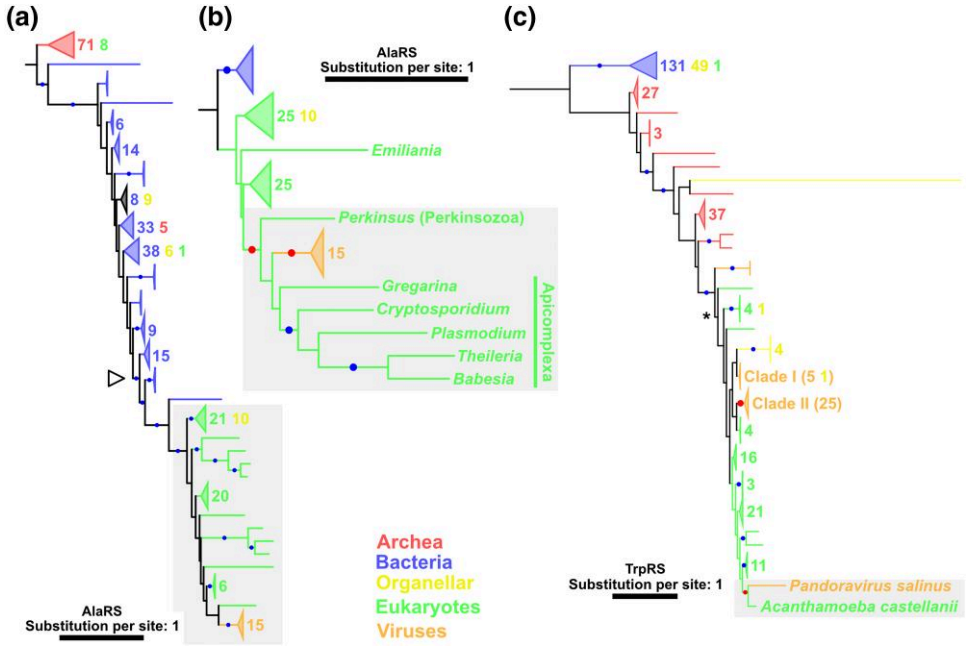

Fig. 4. a) A phylogenetic tree and b) its subsection tree for AlaRS and c) a phylogenetic tree of TrpRS. Red and blue dots indicate significant statistical support (≥80% SH-aLRT and ≥95% UFB). The numbers next to the collapsed clades and their colors indicate the number of sequences and their sources, respectively. Red dots and gray area indicate the nodes and the phylogenetic relationship described in the main text, respectively. a) An empty arrowhead indicates the node from which subsection trees were constructed. b) For eukaryotic sequences, only names of genera are shown. The substitution models were a) LG + R10, b) LG + I + I + R7, and c) LG + R9. The root was decided following the work of Furukawa et al. (2017).

nucleocytoviral, eukaryotic, and organellar sequences. The monophyly of nucleocytovirus sequences and the monophyly of organellar sequences were also statistically supported. Most eukaryotic sequences also formed a statistically supported clade. Only sequences from Metamonada species (*Trichomonas vaginalis*, *Spironucleus salmonicida*, and *Giardia intestinalis*) were located outside of the group of viral, organellar, and other eukaryotic sequences. These results suggest an ancient HGT for CysRS between nucleocytoviruses and eukaryotes before the divergence of most eukaryote lineages.

Recent HGT from Eukaryotes to Nucleocytoviruses (Recent Euk to V)

In addition to the ancient HGTs between eukaryotes and nucleocytoviruses, we identified recent transfers from eukaryotes to nucleocytoviruses in AlaRS, TrpRS, ArgRS, SerRS, MetRS, and ThrRS (Figs. 3 and 4; supplementary figs. S6 and S7, Supplementary Material online). Here, we mainly describe the trees of AlaRS and TrpRS.

In the AlaRS tree, nucleocytovirus AlaRSs formed a statistically supported monophyletic group (Fig. 4a). This viral AlaRS clade forms a statistically supported clade together with eukaryotic AlaRSs, but the position of the viral clade within eukaryotic branches was unclear. To infer the position of the viral clade, we constructed a subsection tree around the viral and eukaryotic clade (Fig. 4b). The viral AlaRSs still formed a statistically supported clade, which was grouped together with AlaRSs from Apicomplexa and Perkinsozoa (two eukaryotic phyla belonging to Alveolata). This grouping of two eukaryotic phyla and a viral clade was statistically supported (Fig. 4b). Because Perkinsozoa and Apicomplexa are closely related phyla within Alveolata (Janouškovec et al. 2015), this clade likely represents the vertical evolution of these eukaryotes. Taken together, we conclude that an HGT occurred for AlaRS from eukaryotes to nucleocytoviruses during the divergence of these eukaryotic phyla.

Several aaRSs identified in the genomes of isolated viruses also showed clear evidence of recent transfer from eukaryotes. One example was found in TrpRS encoded by pandoravirus salinus (Fig. 4c). This TrpRS was grouped together with that from *Acanthamoeba castellanii* with statistical support. Because *A. castellanii* is one of the host organisms of the pandoravirus, this phylogenetic relationship suggests that this viral TrpRS was acquired from its host by HGT. Similar cases were suggested for an ArgRS encoded by orpheovirus and SerRSs encoded by tupanvirus and orpheovirus, although the sources of these HGTs were unclear (supplementary fig. S7, Supplementary Material online).

HGT from Nucleocytoviruses to Eukaryotes (V to Euk)

Putative HGTs from viruses to eukaryotes were suggested for TyrRS, SerRS, AsnRS, and HisRS (Fig. 5; supplementary figs. S7 and S8, Supplementary Material online). A case found in the HisRS tree is described in the Supplementary Material.

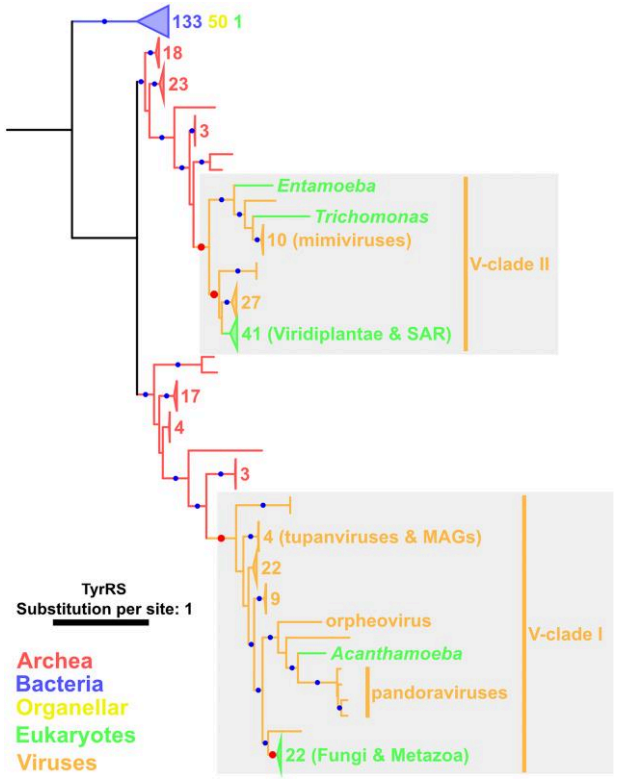

Fig. 5. A phylogenetic tree for TyrRS. Red and blue dots indicate significant statistical support (≥80% SH-aLRT and ≥95% UFB). The numbers next to the collapsed clades and their colors indicate the numbers of sequences and their sources, respectively. Red dots and gray area indicate the nodes and the phylogenetic relationship described in the main text, respectively. For eukaryotic sequences, only names of genera or higher taxonomic levels are shown. The substitution model was LG + R9. The root was decided following the work of Furukawa et al. (2017).

In the TyrRS tree, most eukaryotic sequences were distributed within two statistically supported clades, both of which were included in viral clades (V-clades I and II; statistically supported) (Fig. 5). From the tree topology, V-clades I and II appear to have originated from archaea. The eukaryotic clade in V-clade I includes sequences from Fungi and Metazoa, and that in V-clade II includes sequences from SAR and Viridiplantae. The Fungi and Metazoa clade in V-clade I was grouped with a clade of orpheovirus and pandoraviruses. Sequences from nucleocytoviruses were also present outside of this group. A parsimonious scenario that explains the nested phylogenetic pattern (i.e. a clade of eukaryotes and viruses surrounded further by other viral sequences) involves an ancient HGT of TyrRS from viruses to eukaryotes. The possibility of a scenario involving multiple acquisitions of TyrRS by viruses from ancient eukaryotes cannot be definitively ruled out.

A similar phylogenetic pattern was observed for V-clade II of the TyrRS tree (Fig. 5). Sequences from SAR and Viridiplantae form a clade (statistically unsupported) within V-clade II. The monophyly of the eukaryotic clade and 29 nucleocytovirus sequences was statistically supported, and this group was further surrounded by other viral sequences (including mimiviruses) to form a larger statistically supported clade. This nested phylogenetic pattern suggests that these eukaryotes acquired TyrRS from nucleocytoviruses. Again, we cannot definitively rule out the possibility of multiple acquisitions by nucleocytoviruses from eukaryotes as two eukaryotic sequences (i.e. *Entamoeba* and *Trichomonas*) were placed in the mimivirus clade.

In terms of the direction of HGT occurring from viruses to eukaryotes, there were several more convincing cases. In V-clade I of the TyrRS tree, *A. castellanii* TyrRS was grouped together with pandoraviruses, indicating an HGT from viruses to *A. castellanii* (Fig. 5). Similar virus-to-eukaryote HGT cases were observed in the trees of SerRS (a *Hydra vulgaris* branch in a viral tree; supplementary fig. S7b, Supplementary Material online) and AsnRS (a *Thalassiosira pseudonana* branch embedded in a large viral tree; supplementary fig. S8a, Supplementary Material online). A *Reticulomyxa filosa* branch in a HisRS tree also suggests a similar HGT event, although this scenario was not statistically supported (supplementary fig. S8b, Supplementary Material online).

HGT between Nucleocytoviruses and Prokaryotes (Prok to V)

Nucleocytoviruses exclusively infect eukaryotes, and a majority of detected HGTs were between eukaryotes and nucleocytoviruses. However, our analysis also detected cases of HGTs between viruses and prokaryotes.

A clear case was observed in the tree of ValRS, in which two ValRSs from tupanviruses were grouped together with sequences from *Rickettsia prowazekii* and an organelle of *Ixodes scapularis* (Fig. 6a). This group was further surrounded by bacterial sequences. *R. prowazekii* is an obligate intracellular parasitic bacterium, while *I. scapularis* is a tick species that hosts *Rickettsia* species. These results suggest gene flow from bacteria to nucleocytoviruses and another potential flow from bacteria to ticks. Some *Rickettsia* species are symbionts of ameba, such as *Acanthamoeba* species, which are known as hosts of tupanviruses (Taylor et al. 2012; Abrahão et al. 2018). Therefore, HGT from symbiont bacteria to viruses within ameba cells is also a possible source of gene flow. The ValRS tree shows another possible transfer between bacteria and viruses, although the branching position of the viral clade (containing 12 sequences) was not statistically supported (Fig. 6a).

Another clear case of HGT from bacteria to nucleocytoviruses was inferred in the HisRS tree. One of the three HisRS clades (i.e. clade III), which consists of five viral MAGs, was grouped with bacterial sequences with statistical support, indicating the bacterial origin of this viral sequence clade (supplementary fig. S8b, Supplementary Material online). An ancient bacterial origin for viral sequences was also suggested for clade II of AsnRS, although the true source organisms (bacteria or eukaryotes/organelles) could not be definitively determined (Fig. 3b).

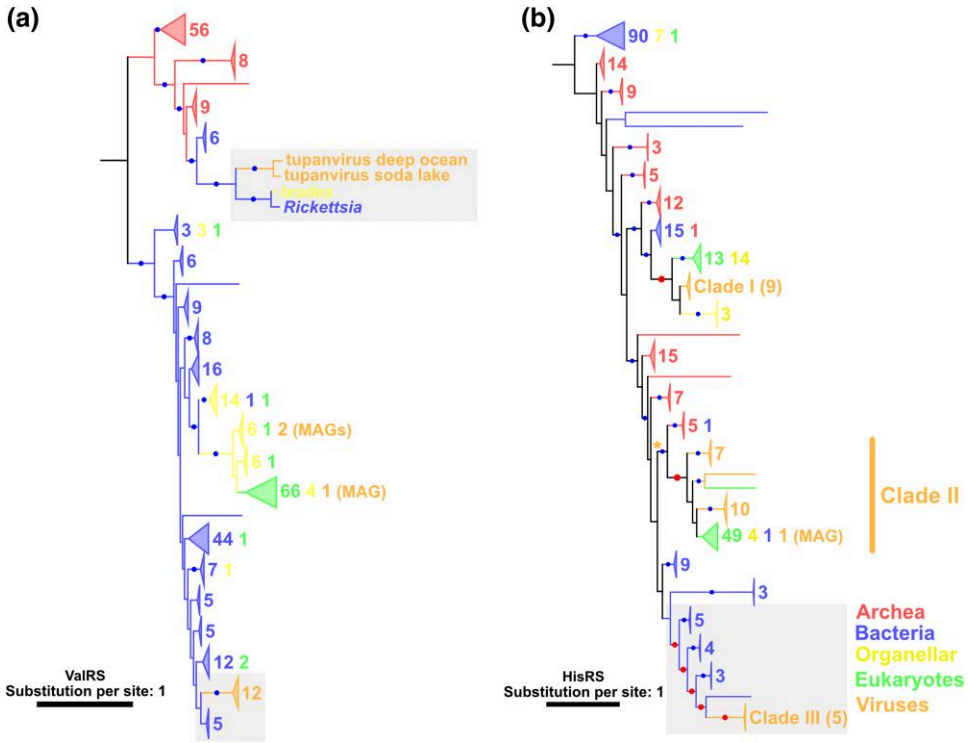

Fig. 6. Phylogenetic trees of a) ValRS and b) HisRS. Red and blue dots indicate nodes with significant statistical support (≥80% SH-aLRT and ≥95% UFB). The numbers next to the collapsed clades and their colors indicate the numbers of sequences and their sources, respectively. Red dots and gray area indicate the nodes and the phylogenetic relationships described in the main text, respectively. b) An asterisk indicates the nodes within which the phylogenetic relationship was further investigated in supplementary fig. S6, Supplementary Material online. The substitution models were a) LG + R10 and b) LG + R10. The root was decided following the work of Furukawa et al. (2017).

We also noticed that the orpheovirus GlyRS sequence was inside an archaeal clade, suggesting a relatively recent HGT from archaea to nucleocytoviruses (Fig. 2a). Additional possible cases of ancient HGT from archaea to viruses are suggested for TyrRS (Fig. 5; as described above).

Evolution of aaRSs within Nucleocytoviruses

Investigation of aaRSs within viral clades further revealed complex evolutionary pathways of these viral aaRSs. Most of the aaRSs were encoded by *Imitervirales* members, which encode 730 out of 780 aaRSs identified in this study, and over half of these aaRSs ($n = 398$) were found in the aaRS-rich clade (Fig. 1). Some aaRSs (ten aaRSs) encoded in this aaRS-rich *Imitervirales* clade appear to have been acquired by the common ancestor of this clade, given the topology of the viral aaRS trees (supplementary fig. S10, Supplementary Material online). In those aaRS trees, clades consisting of sequences from the aaRS-rich *Imitervirales* clade included small numbers of sequences from other *Imitervirales* or viral orders (supplementary fig. S10, Supplementary Material online). These results suggest a scenario in which the common ancestor of the aaRS-rich *Imitervirales* clade acquired these aaRSs from cellular organisms and then spread them to other viruses through virus-to-virus HGTs (hereafter referred to as vHGTs).

In addition to the recent vHGTs, several trees of aaRSs (HisRS, MetRS, ArgRS, and TrpRS) showed the signal of more ancient vHGTs. In these aaRSs, those encoded in the aaRS-rich *Imitervirales* and other *Imitervirales* formed separate clades (Fig. 7). The trees of HisRS and MetRS V-clade I showed statistically supported monophyly of viral aaRSs (Fig. 7a). Within the clades, the separation between the clade for the aaRS-rich *Imitervirales* and that for other *Imitervirales* was statistically supported. In ArgRS and TrpRS, monophyly of the major clade of the aaRS-rich *Imitervirales* was statistically supported and sequences from the other *Imitervirales* were located outside of the clade (Fig. 7b). These results suggest that vHGTs occurred between the common ancestor of the aaRS-rich Imitevirales clades and other *Imitervirales*. We acknowledge that an alternative scenario cannot be definitively ruled out, in which the common ancestor of *Imitervirales* acquired the aaRSs and they were subsequently lost in many lineages outside the aaRS-rich clade over the course of evolution.

Phylogenetic trees of viral AsnRS, IleRS, and LysRS suggested their acquisition by the common ancestor of *Imitervirales* (Fig. 8). Their phylogenetic trees also showed several signals of vHGTs both within *Imitervirales* and with other viral orders. However, a majority of the sequences from the aaRS-rich *Imitervirales* clade form a relatively large clade inside the whole *Imitervirales* clades, which suggests that the common ancestor of *Imitervirales* acquired the aaRSs (Fig. 8). AsnRSs and IleRSs were identified in the genomes of isolates that are closely related to mimiviruses (i.e. megaviruses and moumouviruses), but absent from those of mimiviruses. As megaviruses and moumouviruses formed a clade with other closely related *Imitervirales* (tupanviruses), relatively recent gene loss occurred in the common ancestor of mimiviruses.

As reported in cellular organisms (Wolf et al. 1999), some viruses showed displacements of aaRSs by those derived from other species. In the AspRS tree, the topology suggests that AspRS has been vertically inherited at least within the aaRS-rich *Imitervirales* clade (Fig. 9a). However,

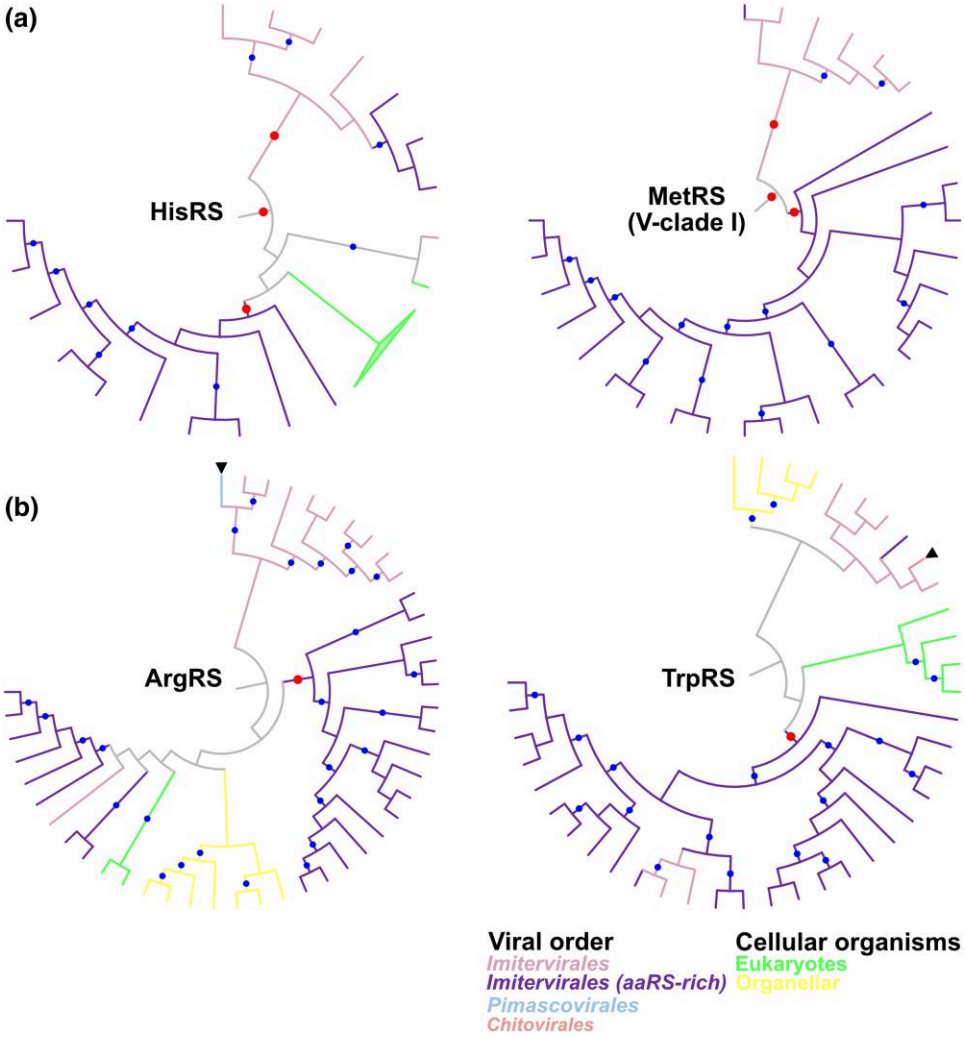

Fig. 7. Phylogenetic trees for viral clades found in a) HisRS and MetRS and b) ArgRS and TrpRS. Names of aaRSs are shown in the center of each tree. These are subtrees extracted from the original trees based on the full set of sequences and the rooting also followed the original topologies. Red and blue dots indicate statistically supported nodes. Red dots indicate the nodes described in the main text. Node and branch colors indicate the order of nucleocytoviruses. Members of the aaRS-rich clade within *Imitervirales* are shown in different colors. Arrowheads indicate putative HGTs between viral orders. Branch length was ignored.

those encoded in tupanviruses were included in a eukaryotic clade with statistical support. These results suggest that tupanvirus AspRSs were displaced by eukaryotic homologs. Another case was found in MetRS encoded by klosneuvirus and its relative, hokovirus. While MetRSs from the aaRS-rich *Imitervirales* clade formed a major monophyletic clade within V-clade I (Fig. 7), MetRSs from klosneuvirus and hokovirus belonged to a phylogenetically distant eukaryotic clade (Fig. 9b). This suggests that klosneuvirus MetRS was displaced by eukaryotic homologs. Putative gene displacements were also identified in ProRS and TyrRS trees (supplementary fig. S11, Supplementary Material online), which we discuss in Supplementary Material.

Discussion

The origin and evolution of viral aaRSs have been debated for around two decades since the discovery of APMV (Raoult et al. 2004; Yutin et al. 2014; Moreira and López-García 2015; Abrahão et al. 2017; Schulz et al. 2017; Abrahão et al. 2018). In the present study, we used an expanded set of nucleocytovirus aaRS sequences, including those from metagenomic data, with a large data set of cellular aaRSs to construct their phylogenetic trees. The resultant trees recovered viral aaRS clades and their relationships to cellular aaRSs with statistical support. Most of these evolutionary relationships were unrecognized in previous studies. On the basis of our updated aaRS phylogeny, recent transfer from eukaryotes to viruses was identified in the trees of AlaRS, MetRS, TrpRS, ArgRS, and SerRS (Figs. 3 and 4; supplementary fig. S7, Supplementary Material online). In contrast, the phylogenetic trees of GlyRS, IleRS, PheRS, AspRS, GluRS, LysRS, and LeuRS suggest scenarios of ancient transfers between proto-eukaryotes and nucleocytoviruses (Fig. 2; supplementary fig. S2, Supplementary Material online). A similar or more ancient origin was suggested in the TyrRS tree, where eukaryotic clades were nested inside the tree of viral sequences (Fig. 5). Regardless of the direction of HGTs (viruses to proto-eukaryotes or multiple acquisitions by viruses from proto-eukaryotes), the origin of TyrRS can date back to the proto-eukaryotic period. In total, the origins of eight aaRSs date back to the time before LECA and four aaRSs predate the divergence of major eukaryotic lineages (Table 1). These ancient aaRSs outnumber those predicted to have ancient origins in a previous study (only two: GluRS

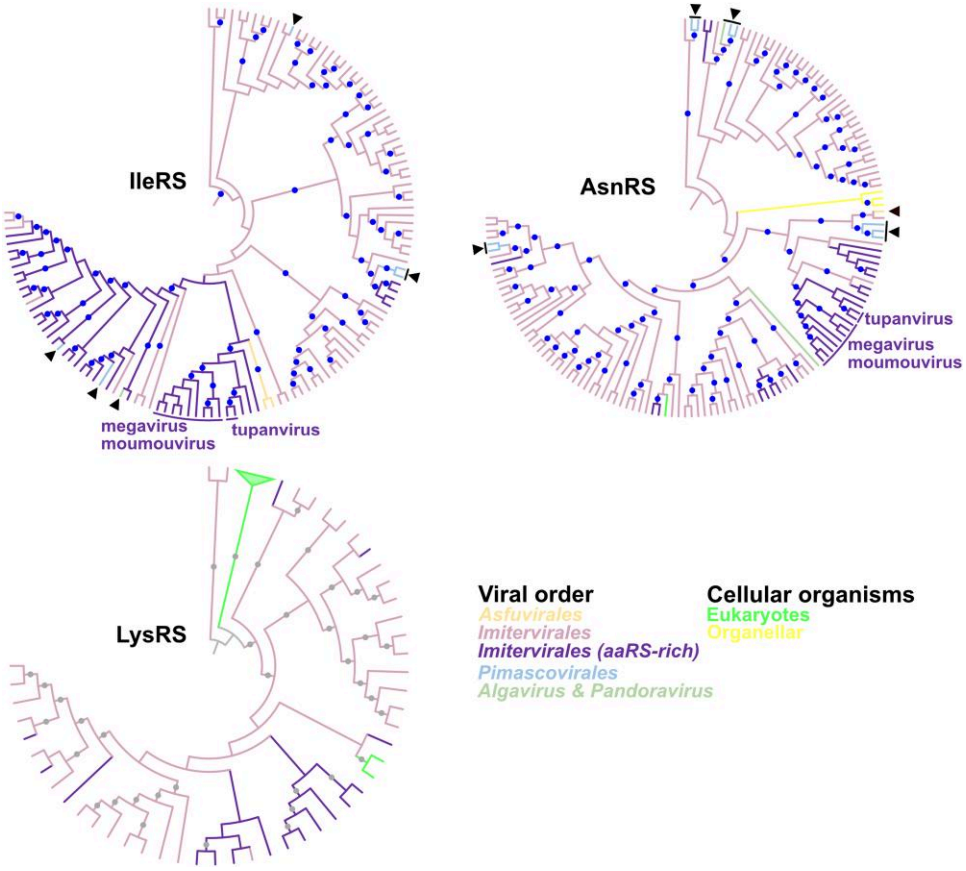

Fig. 8. Phylogenetic trees for viral clades found in IleRS, AsnRS, and LysRS. Names of aaRSs are shown in the center of each tree. These are subtrees extracted from the original trees based on the full set of sequences and the rooting also followed the original topologies. Blue and gray dots indicate nodes statistically supported by UFB and SH-aLRT and those supported by TBE, respectively. Node and branch colors indicate the order of nucleocytoviruses. Members of the aaRS-rich clade within *Imitervirales* are shown in different colors. Aarrowheads indicate putative HGTs between viral orders. Branch length was ignored.

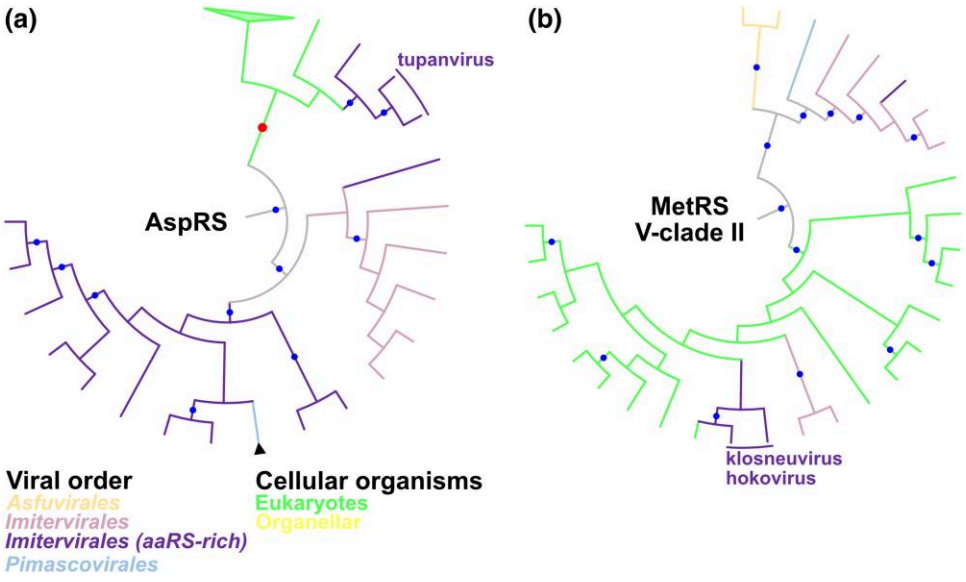

Fig. 9. Phylogenetic trees for viral clades found in a) AspRS and b) MetRS. These are subtrees extracted from the original trees based on the full set of sequences and the rooting also followed the original topologies. Red and blue dots indicate statistically supported nodes. Red dots indicate the nodes described in the main text. Node and branch colors indicate the order of nucleocytoviruses. Members of the aaRS-rich clade within *Imitervirales* are shown in different colors. Arrowheads indicate putative HGTs between viral orders. Branch length was ignored.

and HisRS) (Schulz et al. 2017). Although most HGTs appear to occur between viruses and eukaryotes (including proto-eukaryotes) (Table 1), we also identified statistically supported cases for prokaryote-to-virus HGTs for ValRSs, HisRSs, and GlyRS (Figs. 2 and 6). Taken together, our results establish that nucleocytoviruses acquired their aaRSs from diverse sources of cellular organisms across a wide span of time likely from the proto-eukaryotic period to more recently.

The evolutionary origin of ancestral genomes of nucleocytoviruses is another controversial issue. Some proposed reductive evolution from extinct cellular domains (Raoult et al. 2004; Claverie 2006), while others suggested a model involving accumulation from small ancestral viruses (Yutin et al. 2014; Moreira and López-García 2015; Forterre and Gaïa 2016). The present study showed that some transfers can be traced back to the proto-eukaryotic period, which

indicates that the gene pool of ancestral nucleocytoviruses in this period already included an expanded set of aaRSs (Table 1 and Fig. 2; supplementary fig. S2, Supplementary Material online). Notably, however, some aaRSs (e.g. AlaRS, SerRS, and ArgRS) were embedded in eukaryotic clades and likely to have been recently acquired by nucleocytoviruses, suggesting that the set of aaRSs encoded by ancestral nucleocytoviruses was incomplete. Although reductive evolution can partially account for the current distribution of aaRSs in nucleocytoviruses (as we detected gene losses in mimiviruses; Fig. 8), reduction from the complete set of aaRSs is not supported by the current data. A previous study proposed that the ancestral nucleocytoviruses encoded around 40 genes (Yutin et al. 2014). The expanded set of aaRSs in ancestral nucleocytoviruses as we suggest here does not necessarily contradict this estimate. Instead, it is plausible that this hypothetical small ancestor corresponds to a more ancestral state than the viruses that we consider in this study. We hypothesize that ancestral nucleocytoviruses infecting proto-eukaryotes had already developed complex genomes, as suggested elsewhere (Koonin and Yutin 2010).

Ancient origins of viral aaRSs and their highly biased distribution in *Imitervirales* suggest the ancient order level diversification of *Nucleocytoviricota* that predates the period of LECA. Viruses of the order Imitevirales encode 730 out of 780 nucleocytovirus aaRSs identified in this study (Fig. 1). In particular, the aaRSs with ancient origin predating proto-eukaryotes period were almost exclusively found in the order *Imitervirales*, one of the largest clades in the phylum (Aylward et al. 2021). This distribution pattern indicates that these aaRSs were acquired by a common ancestor of this viral order, which further implies that the common ancestor was already established before the establishment of LECA. Consistently, a previous study inferred that the diversification of viral orders in *Nucleocytoviricota* predates LECA based on the phylogeny of RNA polymerases (Guglielmini et al. 2019). Another recent study identified a distinct group of small viruses that were highly divergent from but related to other large nucleocytoviruses (Yutin et al. 2024), opening the possibility that the root of the viral phylum is deeper than the currently recognized taxonomic framework (Aylward et al. 2021).

In the order *Imitervirales*, we identified a clade particularly enriched in aaRSs (Fig. 1). The phylogenetic distribution of viral aaRSs within the viral clades indicates that this aaRS-rich clade provides its viral aaRSs to the other members of *Imitervirales* through vHGT (Fig. 7; supplementary fig. S10, Supplementary Material online). We also identified HGTs from members of Imitevirales to viruses belonging to other viral orders. These results are consistent with findings in recent studies (Kijima et al. 2021; Da Cunha et al. 2022; Wu et al. 2023) and indicate that vHGT is one of the major drivers behind the spread of viral aaRSs among nucleocytoviruses. The phylogenies of some prevalent aaRSs (i.e. IleRS, AsnRS, and LysRS), in contrast, imply vertical evolution of viral aaRSs from the common ancestor of Imitevirales (Fig. 8). Moreover, we identified the putative displacement of aaRSs (Fig. 9; supplementary fig. S11, Supplementary Material online). Clear cases were observed in AspRS, MetRS, ProRS, and TyrRS of tupanviruses and klosneuviruses. These viruses have a complete or near-complete set of aaRSs, the origin of which has been debated (Schulz et al. 2017; Abrahão et al. 2018). Our phylogenetic trees indicate that their aaRSs are often displaced by eukaryotic homologs, although their ancestors likely inherited aaRSs from the common ancestor of the aaRS-rich *Imitervirales*. These complex evolutionary trajectories involving HGTs, gene losses, and displacements of viral aaRSs may account, at least in part, for the previous disagreements regarding their origins. Nucleocytoviruses and eukaryotes evolve through tight interactions via HGTs in both directions: from eukaryotes to viruses and from viruses to eukaryotes (Guglielmini et al. 2019; Irwin et al. 2022). Our analyses of aaRSs suggest that most cases are HGTs from eukaryotes to viruses. However, we also identified recent HGTs from viruses to eukaryotes, indicating that the HGTs are not limited to the ancient period but still ongoing evolutionary events. Sequences from *A. castellanii*, *H. vulgaris*, *T. pseudonana*, and *R. filosa* are embedded in the viral clades in the trees of TyrRS, SerRS, AsnRS, and HisRS, respectively (Fig. 5; supplementary figs. S7 and S8, Supplementary Material online). *A. castellanii* is a known host of some nucleocytoviruses, suggesting that viral infection sometimes leads to HGTs from viruses to hosts. *H. vulgaris* harbors an endogenous viral region (Filée 2014; Aylward and Moniruzzaman 2021). Recently, giant endogenous viral elements were found in algal species (Moniruzzaman, et al. 2020b) and fungal species (Zhao et al. 2023). Although the functionality of eukaryotic aaRSs derived from viruses is largely unknown, endogenization of viral genomes may account for some of the HGTs from viruses to eukaryotes.

In cellular organisms, aaRSs exceptionally display a high frequency of HGTs despite their essentiality (Doolittle and Handy 1998; Wolf et al. 1999). Nevertheless, their phylogeny is informative to dissect the tree of life (Wolf et al. 1999; Woese et al. 2000; Furukawa et al. 2017). In the present study, we showed that viral aaRSs have similar complex phylogenetic patterns, namely, frequent HGTs, gene losses, and displacements. Meanwhile, by tracing the evolutionary histories of 20 aaRSs, our results suggest that the ancestral nucleocytoviruses already encoded expanded sets of aaRSs in the proto-eukaryotic era. These results demonstrate that viral aaRSs are also informative markers to resolve the evolutionary history of nucleocytoviruses and their relationship with cellular organisms. Previous studies aimed at tracing viral evolution mainly focused on functionally conserved genes (Yutin et al. 2014; Guglielmini et al. 2019; Aylward et al. 2021). Unlike cellular aaRSs, viral aaRSs do not appear to be essential for viruses. The present study established the presence of auxiliary genes in ancient viruses and their usefulness in resolving the deep evolution of nucleocytoviruses. Analysis of other nonessential viral genes may provide further insights into the properties of ancient viral genomes.

Materials and Methods

Sequence Data Set

We constructed nonredundant MAGs from 4,168 nucleocytovirus MAGs constructed in previous studies (Schulz et al. 2020; Moniruzzaman, et al. 2020a; Gaïa et al. 2023). Redundancies among these MAGs were removed as described previously (Kaneko et al. 2021). Briefly, average nucleotide identity (ANI) between MAGs was calculated by dnadiff 1.3 from MUMmer 4.0.0 beta2 (Marçais et al. 2018), and a pair of MAGs that showed ANI greater than 98% with their alignment covering over 25% of the smaller one was clustered together. The largest MAG constructed in *Tara* Oceans Project (Gaïa et al. 2023) was selected as the representative of each cluster. If a cluster contained no MAGs constructed in the *Tara* Oceans Project, the largest MAG was selected as representative. In addition, a data set of 224 reference genomes was obtained from the Global Ocean Eukaryotic Viral database (Gaïa et al. 2023). We used coding regions of sequences predicted in a previous study (Gaïa et al. 2023). Gene annotation was further performed using BLASTP from DIAMOND BLAST v2.0.12.150 (Buchfink et al. 2021) against the UniProtKB database (UniProt Consortium 2021) at an *E*-value threshold of 1*e*−5.

Identification of aaRSs

Genes whose best hit was annotated as an aaRS were considered as putative nucleocytovirus aaRSs. Nonfunctional genes and possible contamination of aaRSs from cellular organisms were excluded from the data set by using phylogenetic trees constructed for each aaRS of nucleocytoviruses and cellular organisms collected from a previous study (Furukawa et al. 2017). Nonfunctional genes were manually identified as short sequences and long-branch sequences of nucleocytovirus. Decontamination of sequences derived from cellular organisms was performed as follows. We first detected nucleocytovirus aaRSs that were not included in clades that comprise multiple nucleocytovirus aaRSs in each tree. If contigs that encode the detected aaRS sequence were from a MAG and contained no nucleocytovirus core gene or gene whose best hit taxonomy was nucleocytovirus, we removed the aaRSs as contamination from cellular organisms.

Phylogenetic Analysis

Nucleocytovirus eight-core genes were identified in the data set by considering the DIAMOND BLASTP hits that were registered in NCVOG (Yutin et al. 2009). Multiple sequence alignments were performed by MAFFT v7.487 (Katoh and Standley 2013) with the E-INS-i algorithm, which is suitable for the alignment of sequences in which the number of domains is unknown. Subsequently, the sites with gaps in more than 75% of aligned sequences were removed by using trimAl v1.4.rev15 (Capella-Gutiérrez et al. 2009). Phylogenetic trees were constructed within the maximum likelihood framework by IQ-TREE 2.1.3 (Minh et al. 2020). The amino acid substitution model was selected by Model Finder (Kalyaanamoorthy et al. 2017). Bootstrap values were computed by the SH-aLRT method (Guindon et al. 2010) and the UFB method (Hoang et al. 2018). Clades with >80% SH-aLRT and >95% UFB were considered to be statistically supported. The tree in Fig. 1 was visualized using anvi'o v. 8 (Eren et al. 2021), and other trees were visualized by iTOL v6 (Letunic and Bork 2021). The root of each tree was decided following a previous study (Furukawa et al. 2017). In the previous study, the root was set at a branch that separates major bacterial and archaeal clades of each aaRS. This root was proposed by phylogenetic analysis of duplicated genes that are well conserved in the three domains of life, including multiple aaRSs (Iwabe et al. 1989; Brown and Doolittle 1995; Brown et al. 1997).

The origin of nucleocytovirus aaRS was inferred using phylogenetic trees constructed as described above. We ignored the clades comprising fewer than three nucleocytovirus sequences, to reduce the contamination from nucleocytovirus MAGs contaminated by contigs of cellular organisms. We also constructed phylogenetic trees from subsections of data used for tree reconstruction above. These subsections contain a nucleocytovirus clade and its neighboring cellular branches. In addition to SH-aLRT and UFB, transfer bootstrap expectation (TBE) (Lemoine et al. 2018) was calculated. While conventional bootstrap approaches count exactly the same branches in bootstrap trees, TBE takes into account similar branches by using transfer distances (Lemoine et al. 2018). The use of TBE support facilitates inferences about the presence of unstable taxa (single taxa that tend to move in and out of clades) and is considered to be particularly useful for deep branches and large data sets. Here, clades supported with >70% TBE were considered to be statistically supported.

For seven aaRSs (GlyRS, IleRS, PheRS, AspRS, GluRS, LeuRS, and LysRS), whose evolutionary scenario was inferred as proto-Euk&V, we performed additional analysis to examine the robustness of the scenario. From the multiple sequence alignments of these aaRSs, the sites with gaps in more than 50% of aligned sequences were removed by trimAl and phylogenetic trees were built by IQ-TREE. We also built constrained trees of these aaRSs with 15 topologies and performed the AU test to calculate *P* values by IQ-TREE. To build constrained trees, eukaryotic clades were separated into three clades at the top two bipartitions within the major eukaryotic clades. In this process, bipartitions involving singletons or clades with two sequences were skipped, and these sequences were excluded in the phylogenetic analyses. For GlyRS, LeuRS, and LysRS, the subsets of sequences were examined.

Supplementary Material

Supplementary material is available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank the *Tara* Oceans consortium and the people and sponsors who supported *Tara* Oceans. *Tara* Oceans

(including both the *Tara* Oceans and *Tara* Oceans Polar Circle expeditions) would not exist without the leadership of the *Tara* Expeditions Foundation and the continuous support of 23 institutes (https://oceans.taraexpeditions.org). This article is contribution number 153 of *Tara* Oceans. We also thank Edanz (https://jp.edanz.com/ac) for editing a draft of this manuscript. Computation time was provided by the supercomputer system, Institute for Chemical Research, Kyoto University.

Funding

This work was supported in part by the Japan Society for the Promotion of Science KAKENHI (22H00384), the Research Unit for Development of Global Sustainability, the Kyoto University Research Coordination Alliance, and the International Collaborative Research Program of the Institute for Chemical Research, Kyoto University (202332, 2022-26, 2021-29, and 2020-28).

Conflict of Interest

The authors declare that there are no conflicts of interest in this research.

Data Availability

The phylogenetic tree data generated in this study are also available from GenomeNet (https://www.genome.jp/ftp/db/community/tara/aaRS/).

References

Abrahão JS, Araújo R, Colson P, La Scola B. The analysis of translation-related gene set boosts debates around origin and evolution of mimiviruses. *PLoS Genet*. 2017:**13**(2):e1006532. https://doi.org/10.1371/journal.pgen.1006532.

Abrahão J, Silva L, Silva LS, Khalil JYB, Rodrigues R, Arantes T, Assis F, Boratto P, Andrade M, Kroon EG, *et al*. Tailed giant tupanvirus possesses the most complete translational apparatus of the known virosphere. *Nat Commun*. 2018:**9**(1):749. https://doi.org/10.1038/s41467-018-03168-1.

Arslan D, Legendre M, Seltzer V, Abergel C, Claverie J-M. Distant mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. *Proc Natl Acad Sci U S A*. 2011:**108**(42):17486–17491. https://doi.org/10.1073/pnas.1110889108.

Aylward FO, Moniruzzaman M. ViralRecall—a flexible command-line tool for the detection of giant virus signatures in 'omic data. *Viruses*. 2021:**13**(2):150. https://doi.org/10.3390/v13020150.

Aylward FO, Moniruzzaman M, Ha AD, Koonin EV. A phylogenomic framework for charting the diversity and evolution of giant viruses. *PLoS Biol*. 2021:**19**(10):e3001430. https://doi.org/10.1371/journal.pbio.3001430.

Blanc-Mathieu R, Dahle H, Hofgaard A, Brandt D, Ban H, Kalinowski J, Ogata H, Sandaa R-A. A persistent giant algal virus, with a unique morphology, encodes an unprecedented number of genes involved in energy metabolism. *J Virol*. 2021:**95**(8):e02446-20. https://doi.org/10.1128/JVI.02446-20.

Boughalmi M, Saadi H, Pagnier I, Colson P, Fournous G, Raoult D, La Scola B. High-throughput isolation of giant viruses of the Mimiviridae and Marseilleviridae families in the Tunisian environment. *Environ Microbiol*. 2013:**15**(7):2000–2007. https://doi.org/10.1111/1462-2920.12068.

Boyer M, Madoui M-A, Gimenez G, La Scola B, Raoult D. Phylogenetic and phyletic studies of informational genes in genomes highlight existence of a 4 domain of life including giant viruses. *PLoS One*. 2010:**5**(12):e15530. https://doi.org/10.1371/journal.pone.0015530.

Boyer M, Yutin N, Pagnier I, Barrassi L, Fournous G, Espinosa L, Robert C, Azza S, Sun S, Rossmann MG, *et al*. Giant *Marseillevirus* highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. *Proc Natl Acad Sci U S A*. 2009:**106**(51):21848–21853. https://doi.org/10.1073/pnas.0911354106.

Brown JR, Doolittle WF. Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. *Proc Natl Acad Sci U S A*. 1995:**92**(7):2441–2445. https://doi.org/10.1073/pnas.92.7.2441.

Brown JR, Robb FT, Weiss R, Doolittle WF. Evidence for the early divergence of tryptophanyl- and tyrosyl-tRNA synthetases. *J Mol Evol*. 1997:**45**(1):9–16. https://doi.org/10.1007/PL00006206.

Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods*. 2021:**18**(4):366–368. https://doi.org/10.1038/s41592-021-01101-x.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. Trimal: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*. 2009:**25**(15):1972–1973. https://doi.org/10.1093/bioinformatics/btp348.

Claverie J-M. Viruses take center stage in cellular evolution. *Genome Biol*. 2006:**7**(6):110. https://doi.org/10.1186/gb-2006-7-6-110.

Da Cunha V, Gaia M, Ogata H, Jaillon O, Delmont TO, Forterre P. Giant viruses encode actin-related proteins. *Mol Biol Evol*. 2022:**39**(2):msac022. https://doi.org/10.1093/molbev/msac022.

Doolittle RF, Handy J. Evolutionary anomalies among the aminoacyl-tRNA synthetases. *Curr Opin Genet Dev*. 1998:**8**(6):630–636. https://doi.org/10.1016/S0959-437X(98)80030-0.

Eren AM, Kiefl E, Shaiber A, Veseli I, Miller SE, Schechter MS, Fink I, Pan JN, Yousef M, Fogarty EC, *et al*. Community-led, integrated, reproducible multi-omics with anvi'o. *Nat Microbiol*. 2021:**6**(1):3–6. https://doi.org/10.1038/s41564-020-00834-3.

Filée J. Multiple occurrences of giant virus core genes acquired by eukaryotic genomes: the visible part of the iceberg? *Virology*. 2014:**466–467**:53–59. https://doi.org/10.1016/j.virol.2014.06.004.

Forterre P, Gaïa M. Giant viruses and the origin of modern eukaryotes. *Curr Opin Microbiol*. 2016:**31**:44–49. https://doi.org/10.1016/j.mib.2016.02.001.

Furukawa R, Nakagawa M, Kuroyanagi T, Yokobori S-I, Yamagishi A. Quest for ancestors of eukaryal cells based on phylogenetic analyses of aminoacyl-tRNA synthetases. *J Mol Evol*. 2017:**84**(1):51–66. https://doi.org/10.1007/s00239-016-9768-2.

Gaïa M, Meng L, Pelletier E, Forterre P, Vanni C, Fernandez-Guerra A, Jaillon O, Wincker P, Ogata H, Krupovic M, *et al*. Mirusviruses link herpesviruses to giant viruses. *Nature*. 2023:**616**(7958):783–789. https://doi.org/10.1038/s41586-023-05962-4.

Guglielmini J, Gaia M, Da Cunha V, Criscuolo A, Krupovic M, Forterre P. Viral origin of eukaryotic type IIA DNA topoisomerases. *Virus Evol*. 2022:**8**(2):veac097. https://doi.org/10.1093/ve/veac097.

Guglielmini J, Woo AC, Krupovic M, Forterre P, Gaia M. Diversification of giant and large eukaryotic dsDNA viruses predated the origin of modern eukaryotes. *Proc Natl Acad Sci U S A*. 2019:**116**(39):19585–19592. https://doi.org/10.1073/pnas.1912006116.

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol*. 2010:**59**(3):307–321. https://doi.org/10.1093/sysbio/syq010.

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: improving the ultrafast bootstrap approximation. *Mol Biol Evol*. 2018:**35**(2):518–522. https://doi.org/10.1093/molbev/msx281.

Irwin NAT, Pittis AA, Richards TA, Keeling PJ. Systematic evaluation of horizontal gene transfer between eukaryotes and viruses. *Nat Microbiol*. 2022:**7**(2):327–336. https://doi.org/10.1038/s41564-021-01026-3.

Iwabe N, Kuma K, Hasegawa M, Osawa S, Miyata T. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc Natl Acad Sci U S A*. 1989:**86**(23):9355–9359. https://doi.org/10.1073/pnas.86.23.9355.

Iyer LM, Balaji S, Koonin EV, Aravind L. Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. *Virus Res*. 2006:**117**(1):156–184. https://doi.org/10.1016/j.virusres.2006.01.009.

Janouškovec J, Tikhonenkov DV, Burki F, Howe AT, Kolísko M, Mylnikov AP, Keeling PJ. Factors mediating plastid dependency and the origins of parasitism in apicomplexans and their close relatives. *Proc Natl Acad Sci U S A*. 2015:**112**(33):10200–10207. https://doi.org/10.1073/pnas.1423790112.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods*. 2017:**14**(6):587–589. https://doi.org/10.1038/nmeth.4285.

Kaneko H, Blanc-Mathieu R, Endo H, Chaffron S, Delmont TO, Gaia M, Henry N, Hernández-Velázquez R, Nguyen CH, Mamitsuka H, *et al.* Eukaryotic virus composition can predict the efficiency of carbon export in the global ocean. *iScience*. 2021:**24**(1):102002. https://doi.org/10.1016/j.isci.2020.102002.

Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 2013:**30**(4):772–780. https://doi.org/10.1093/molbev/mst010.

Kijima S, Delmont TO, Miyazaki U, Gaia M, Endo H, Ogata H. Discovery of viral myosin genes with complex evolutionary history within plankton. *Front Microbiol*. 2021:**12**:683294. https://doi.org/10.3389/fmicb.2021.683294.

Koonin EV, Yutin N. Origin and evolution of eukaryotic large nucleocytoplasmic DNA viruses. *Intervirology*. 2010:**53**(5):284–292. https://doi.org/10.1159/000312913.

Legendre M, Fabre E, Poirot O, Jeudy S, Lartigue A, Alempic J-M, Beucher L, Philippe N, Bertaux L, Christo-Foroux E, *et al.* Diversity and evolution of the emerging Pandoraviridae family. *Nat Commun*. 2018:**9**(1):2285. https://doi.org/10.1038/s41467-018-04698-4.

Lemoine F, Domelevo Entfellner J-B, Wilkinson E, Correia D, Dávila Felipe M, De Oliveira T, Gascuel O. Renewing Felsenstein's phylogenetic bootstrap in the era of big data. *Nature*. 2018:**556**(7702):452–456. https://doi.org/10.1038/s41586-018-0043-0.

Letunic I, Bork P. Interactive Tree of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res*. 2021:**49**(W1):W293–W296. https://doi.org/10.1093/nar/gkab301.

Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: a fast and versatile genome alignment system. *PLoS Comput Biol*. 2018:**14**(1):e1005944. https://doi.org/10.1371/journal.pcbi.1005944.

Mihara T, Koyano H, Hingamp P, Grimsley N, Goto S, Ogata H. Taxon richness of "Megaviridae" exceeds those of bacteria and archaea in the ocean. *Microbes Environ*. 2018:**33**(2):162–171. https://doi.org/10.1264/jsme2.ME17203.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol*. 2020:**37**(5):1530–1534. https://doi.org/10.1093/molbev/msaa015.

Moniruzzaman M, Martinez-Gutierrez CA, Weinheimer AR, Aylward FO. Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. *Nat Commun*. 2020a:**11**(1):1710. https://doi.org/10.1038/s41467-020-15507-2.

Moniruzzaman M, Weinheimer AR, Martinez-Gutierrez CA, Aylward FO. Widespread endogenization of giant viruses shapes genomes of green algae. *Nature*. 2020b:**588**(7836):141–145. https://doi.org/10.1038/s41586-020-2924-2.

Moreira D, López-García P. Evolution of viruses and cells: do we need a fourth domain of life to explain the origin of eukaryotes? *Philos Trans R Soc Lond B Biol Sci*. 2015:**370**(1678):20140327. https://doi.org/10.1098/rstb.2014.0327.

Philippe N, Legendre M, Doutre G, Couté Y, Poirot O, Lescot M, Arslan D, Seltzer V, Bertaux L, Bruley C, *et al.* Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science*. 2013:**341**(6143):281–286. https://doi.org/10.1126/science.1239181.

Raoult D, Audic S, Robert C, Abergel C, Renesto P, Ogata H, La Scola B, Suzan M, Claverie J-M. The 1.2-megabase genome sequence of mimivirus. *Science*. 2004:**306**(5700):1344–1350. https://doi.org/10.1126/science.1101485.

Schulz F, Roux S, Paez-Espino D, Jungbluth S, Walsh DA, Denef VJ, McMahon KD, Konstantinidis KT, Eloe-Fadrosh EA, Kyrpides NC, *et al.* Giant virus diversity and host interactions through global metagenomics. *Nature*. 2020:**578**(7795):432–436. https://doi.org/10.1038/s41586-020-1957-x.

Schulz F, Yutin N, Ivanova NN, Ortega DR, Lee TK, Vierheilig J, Daims H, Horn M, Wagner M, Jensen GJ, *et al.* Giant viruses with an expanded complement of translation system components. *Science*. 2017:**356**(6333):82–85. https://doi.org/10.1126/science.aal4657.

Schvarcz CR, Steward GF. A giant virus infecting green algae encodes key fermentation genes. *Virology*. 2018:**518**:423–433. https://doi.org/10.1016/j.virol.2018.03.010.

Shimodaira H. An approximately unbiased test of phylogenetic tree selection. *Syst Biol*. 2002:**51**(3):492–508. https://doi.org/10.1080/10635150290069913.

Taylor M, Mediannikov O, Raoult D, Greub G. Endosymbiotic bacteria associated with nematodes, ticks and amoebae. *FEMS Immunol Med Microbiol*. 2012:**64**(1):21–31. https://doi.org/10.1111/j.1574-695X.2011.00916.x.

The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res*. 2021:**49**(D1):D480–D489. https://doi.org/10.1093/nar/gkaa1100.

Williams TA, Embley TM, Heinz E. Informational gene phylogenies do not support a fourth domain of life for nucleocytoplasmic large DNA viruses. *PLoS One*. 2011:**6**(6):e21080. https://doi.org/10.1371/journal.pone.0021080.

Woese CR, Olsen GJ, Ibba M, Söll D. Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. *Microbiol Mol Biol Rev*. 2000:**64**(1):202–236. https://doi.org/10.1128/MMBR.64.1.202-236.2000.

Wolf YI, Aravind L, Grishin NV, Koonin EV. Evolution of aminoacyl-tRNA synthetases–analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. *Genome Res*. 1999:**9**(8):689–710. https://doi.org/10.1101/gr.9.8.689.

Wu J, Meng L, Gaïa M, Hikida H, Okazaki Y, Endo H, Ogata H. Gene transfer among viruses substantially contributes to gene gain of giant viruses. *Mol Biol Evol*. 2024.

Yoshikawa G, Blanc-Mathieu R, Song C, Kayama Y, Mochizuki T, Murata K, Ogata H, Takemura M. Medusavirus, a novel large DNA virus discovered from hot spring water. *J Virol*. 2019:**93**(8):e02130-18. https://doi.org/10.1128/JVI.02130-18.

Yutin N, Mutz P, Krupovic M, Koonin EV. Mriyaviruses: small relatives of giant viruses. *mBio*. 2024:**15**(7):e01035-24. https://doi.org/10.1128/mbio.01035-24.

Yutin N, Wolf YI, Koonin EV. Origin of giant viruses from smaller DNA viruses not from a fourth domain of cellular life. *Virology*. 2014:**466–467**:38–52. https://doi.org/10.1016/j.virol.2014.06.032.

Yutin N, Wolf YI, Raoult D, Koonin EV. Eukaryotic large nucleocytoplasmic DNA viruses: clusters of orthologous genes and reconstruction of viral genome evolution. *Virol J*. 2009:**6**(1):223. https://doi.org/10.1186/1743-422X-6-223.

Zhao H, Zhang R, Wu J, Meng L, Okazaki Y, Hikida H, Ogata H. A 1.5-mb continuous endogenous viral region in the arbuscular mycorrhizal fungus rhizophagus irregularis. *Virus Evol*. 2023:**9**(2):vead064. doi:10.1093/ve/vead064.